Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Chapter 1

OUTLIER DETECTION

Irad Ben-Gal
Department of Industrial Engineering
Tel-Aviv University
Ramat-Aviv, Tel-Aviv 69978, Israel.
bengal@eng.tau.ac.il

Abstract Outlier detection is a primary step in many data-mining applications. We present several methods for outlier detection, while distinguishing between univariate vs. multivariate techniques and parametric vs. nonparametric procedures. In presence of outliers, special attention should be taken to assure the robustness of the used estimators. Outlier detection for data mining is often based on distance measures, clustering and spatial methods.

Keywords: Outliers, Distance measures, Statistical Process Control, Spatial data

1. Introduction: Motivation, Definitions and Applications

In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis (Williams *et al*., 2002; Liu *et al*., 2004).

An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (Hawkins, 1980) defines an outlier *as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.* Barnet and Lewis (Barnett and Lewis, 1994) indicate that *an outlying observation, or outlier, is one that appears to*

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

deviate markedly from other members of the sample in which it occurs, similarly, Johnson (Johnson, 1992) defines an outlier *as an observation in a data set which appears to be inconsistent with the remainder of that set of data*. Other case-specific definitions are given below.

Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks (Hawkins, 1980; Barnett and Lewis, 1994; Ruts and Rousseeuw, 1996; Fawcett and Provost, 1997; Johnson *et al*., 1998; Penny and Jolliffe, 2001; Acuna and Rodriguez, 2004; Lu *et al*., 2003).

2. Taxonomy of Outlier Detection Methods

Outlier detection methods can be divided between *univariate methods*, proposed in earlier works in this field, and *multivariate methods* that usually form most of the current body of research. Another fundamental taxonomy of outlier detection methods is between *parametric* (*statistical*) methods and *non-parametric* methods that are model-free (e.g., see (Williams *et al*., 2002)). Statistical parametric methods either assume a known underlying distribution of the observations (e.g., (Hawkins, 1980; Rousseeuw and Leory, 1987; Barnett and Lewis, 1994)) or, at least, they are based on statistical estimates of unknown distribution parameters (Hadi, 1992; Caussinus and Roiz, 1990). These methods flag as outliers those observations that deviate from the model assumptions. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution (Papadimitriou *et al*., 2002).

Within the class of non-parametric outlier detection methods one can set apart the data-mining methods, also called *distance-based methods*. These methods are usually based on local distance measures and are capable of handling large databases (Knorr and Ng, 1997; Knorr and Ng, 1998; Fawcett and Provost, 1997; Williams and Huang, 1997; Mouchel and Schonlau, 1998; Knorr *et al*., 2000; Knorr *et al*., 2001; Jin *et al*., 2001; Breunig *et al*., 2000; Williams *et al*., 2002; Hawkins *et al*., 2002; Bay and Schwabacher, 2003). Another class of outlier detection methods is founded on *clustering techniques*, where a cluster of small sizes can be considered as clustered outliers (Kaufman and Rousseeuw, 1990; Ng and Han, 1994; Ramaswamy *et al*., 2000; Barbara and Chen, 2000; Shekhar and Chawla, 2002; Shekhar and Lu, 2001; Shekhar and Lu, 2002; Acuna and Rodriguez, 2004). Hu and Sung (Hu and Sung, 2003), whom proposed a method to identify both high and low density pattern clustering, further partition this class to *hard classifiers* and *soft classifiers*. The former partition the data into two non-overlapping sets: outliers and non-

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

outliers. The latter offers a ranking by assigning each datum an outlier classification factor reflecting its degree of outlyingness. Another related class of methods consists of detection techniques for *spatial outliers*. These methods search for extreme observations or local instabilities with respect to neighboring values, although these observations may not be significantly different from the entire population (Schiffman *et al*., 1981; Ng and Han, 1994; Shekhar and Chawla, 2002; Shekhar and Lu, 2001; Shekhar and Lu, 2002; Lu *et al*., 2003).

Some of the above-mentioned classes are further discussed bellow. Other categorizations of outlier detection methods can be found in (Barnett and Lewis, 1994; Papadimitriou *et al*., 2002; Acuna and Rodriguez, 2004; Hu and Sung, 2003).

3. Univariate Statistical Methods

Most of the earliest univariate methods for outlier detection rely on the assumption of an underlying known distribution of the data, which is assumed to be identically and independently distributed (i.i.d.). Moreover, many discordance tests for detecting univariate outliers further assume that the distribution parameters and the type of expected outliers are also known (Barnett and Lewis, 1994). Needless to say, in real world data-mining applications these assumptions are often violated.

A central assumption in statistical-based methods for outlier detection, is a generating model that allows a small number of observations to be randomly sampled from distributions G_1,..., G_k, differing from the target distribution F, which is often taken to be a normal distribution $N\left(\mu,\sigma^2\right)$ (see (Ferguson, 1961; David, 1979; Barnett and Lewis, 1994; Gather, 1989; Davies and Gather, 1993)). The outlier identification problem is then translated to the problem of identifying those observations that lie in a so-called *outlier region*. This leads to the following definition (Davies and Gather, 1993):

For any *confidence coefficient* α, $0 < \alpha < 1$, the α-outlier region of the $N\left(\mu,\sigma^2\right)$ distribution is defined by

$$\text{out}\left(\alpha,\mu,\sigma^2\right) = \left\{x : |x-\mu| > z_{1-\alpha/2}\sigma\right\}, \tag{1.1}$$

where z_q is the q quintile of the $N(0,1)$. A number x is an α-outlier with respect to F if $x \in$ out $\left(\alpha,\mu,\sigma^2\right)$. Although traditionally the normal distribution has been used as the target distribution, this definition can be easily extended to any unimodal symmetric distribution with positive density function, including the multivariate case.

Note that the outlier definition does not identify which of the observations are contaminated, i.e., resulting from distributions G_1,..., G_k , but rather it indicates those observations that lie in the outlier region.

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

3.1 Single-step vs. Sequential Procedures

Davis and Gather (Davies and Gather, 1993) make an important distinction between *single-step* and *sequential* procedures for outlier detection. Single-step procedures identify all outliers at once as opposed to successive elimination or addition of datum. In the sequential procedures, at each step, one observation is tested for being an outlier.

With respect to Equation 1.1, a common rule for finding the outlier region in a single-step identifier is given by

$$\text{out}\left(\alpha_n, \hat{\mu}_n, \hat{\sigma}_n^2\right) = \left\{x : |x - \hat{\mu}_n| > g\left(n, \alpha_n\right)\hat{\sigma}_n\right\}, \tag{1.2}$$

where n is the size of the sample; $\hat{\mu}_n$ and $\hat{\sigma}_n$ are the estimated mean and standard deviation of the target distribution based on the sample; α_n denotes the confidence coefficient following the correction for multiple comparison tests; and $g\left(n, \alpha_n\right)$ defines the limits (critical number of standard deviations) of the outlier regions.

Traditionally, $\hat{\mu}_n, \hat{\sigma}_n$ are estimated respectively by the sample mean, $\bar{x}_n$, and the sample standard deviation, S_n. Since these estimates are highly affected by the presence of outliers, many procedures often replace them by other, more robust, estimates that are discussed in Section . The multiple-comparison correction is used when several statistical tests are being performed simultaneously. While a given α-value may be appropriate to decide whether a single observation lies in the outlier region (i.e., a single comparison), this is not the case for a set of several comparisons. In order to avoid spurious positives, the α-value needs to be lowered to account for the number of performed comparisons. The simplest and most conservative approach is the Bonferroni's correction, which sets the α-value for the entire set of n comparisons equal to α, by taking the α-value for each comparison equal to α/n. Another popular and simple correction uses $\alpha_n = 1 - (1 - \alpha)^{1/n}$. Note that the traditional Bonferroni's method is "quasi-optimal" when the observations are independent, which is in most cases unrealistic. The critical value $g\left(n, \alpha_n\right)$ is often specified by numerical procedures, such as Monte Carlo simulations for different sample sizes (e.g., (Davies and Gather, 1993)).

3.2 Inward and Outward Procedures

Sequential identifiers can be further classified to *inward* and *outward* procedures. In inward testing, or *forward selection* methods, at each step of the procedure the "most extreme observation", i.e., the one with the largest outlyingness measure, is tested for being an outlier. If it is declared as an outlier, it is deleted from the dataset and the procedure is repeated. If it is declared as a non-outlying observation, the procedure terminates. Some classical examples

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

for inward procedures can be found in (Hawkins, 1980; Barnett and Lewis, 1994).

In outward testing procedures, the sample of observations is first reduced to a smaller sample (e.g., by a factor of two), while the removed observations are kept in a reservoir. The statistics are calculated on the basis of the reduced sample and then the removed observations in the reservoir are tested in reverse order to indicate whether they are outliers. If an observation is declared as an outlier, it is deleted from the reservoir. If an observation is declared as a non-outlying observation, it is deleted from the reservoir, added to the reduced sample, the statistics are recalculated and the procedure repeats itself with a new observation. The outward testing procedure is terminated when no more observations are left in the reservoir. Some classical examples for inward procedures can be found in (Rosner, 1975; Hawkins, 1980; Barnett and Lewis, 1994).

The classification to inward and outward procedures also applies to multivariate outlier detection methods.

3.3 Univariate Robust Measures

Traditionally, the sample mean and the sample variance give good estimation for data location and data shape if it is not contaminated by outliers. When the database is contaminated, those parameters may deviate and significantly affect the outlier-detection performance.

Hampel (Hampel, 1971; Hampel, 1974) introduced the concept of the *breakdown point*, as a measure for the robustness of an estimator against outliers. The breakdown point is defined as the smallest percentage of outliers that can cause an estimator to take arbitrary large values. Thus, the larger breakdown point an estimator has, the more robust it is. For example, the sample mean has a breakdown point of $1/n$since a single large observation can make the sample mean and variance cross any bound. Accordingly, Hampel suggested the median and the median absolute deviation (MAD) as robust estimates of the location and the spread. The Hampel identifier is often found to be practically very effective (Perarson, 2002; Liu *et al.*, 2004). Another earlier work that addressed the problem of robust estimators was proposed by Tukey (Tukey, 1977). Tukey introduced the Boxplot as a graphical display on which outliers can be indicated. The Boxplot, which is being extensively used up to date, is based on the distribution quadrants. The first and third quadrants, Q_1 and Q_3, are used to obtain the robust measures for the mean, $\hat{\mu}_n = (Q_1 + Q_3)/2$, and the standard deviation, $\hat{\sigma}_n = Q_3 - Q_1$. Another popular solution to obtain robust measures is to replace the mean by the median and compute the standard deviation based on (1–α) percents of the data points, where typically α¡5%.

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Liu et al. (Liu *et al*., 2004) proposed an outlier-resistant data filter-cleaner based on the earlier work of Martin and Thomson (Martin and Thomson, 1982). The proposed data filter-cleaner includes an on-line outlier-resistant estimate of the process model and combines it with a modified Kalman filter to detect and "clean" outliers. The proposed method does not require an apriori knowledge of the process model. It detects and replaces outliers on-line while preserving all other information in the data. The authors demonstrated that the proposed filter-cleaner is efficient in outlier detection and data cleaning for autocorrelated and even non-stationary process data.

3.4 Statistical Process Control (SPC)

The field of Statistical Process Control (SPC) is closely-related to univariate outlier detection methods. It considers the case where the univariable stream of measures represents a stochastic process, and the detection of the outlier is required online. SPC methods are being applied for more than half a century and were extensively investigated in statistics literature.

Ben-Gal et al. (Gal *et al*., 2003) categorize SPC methods by two major criteria: i) methods for *independent* data versus methods for *dependent* data; and ii) methods that are *model-specific*, versus methods that are *model-generic*. Model specific methods require a-priori assumptions on the process characteristics, usually defined by an underlying analytical distribution or a closed-form expression. Model-generic methods try to estimate the underlying model with minimum a-priori assumptions.

Traditional SPC methods, such as Shewhart, Cumulative Sum (CUSUM) and Exponential Weighted Moving Average (EWMA) are *model-specific* for *independent data*. Note that these methods are extensively implemented in industry, although the independence assumptions are frequently violated in practice.

The majority of *model-specific* methods for *dependent data* are based on time-series. Often, the underlying principle of these methods is as follows: find a time series model that can best capture the autocorrelation process, use this model to filter the data, and then apply traditional SPC schemes to the stream of residuals. In particular, the ARIMA (Auto Regressive Integrated Moving Average) family of models is widely implemented for the estimation and filtering of process autocorrelation. Under certain assumptions, the residuals of the ARIMA model are independent and approximately normally distributed, to which traditional SPC can be applied. Furthermore, it is commonly conceived that ARIMA models, mostly the simple ones such as AR(see Equation 1.1), can effectively describe a wide variety of industry processes (Box, 1976; Apley and Shi, 1999).

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Model-specific methods for dependent data can be further partitioned to *parameter-dependent* methods that require explicit estimation of the model parameters (e.g., (Alwan and Roberts, 1988; Wardell *et al*., 1994; Lu and Reynolds, 1999; Runger and Willemain, 1995; Apley and Shi, 1999)), and to *parameter-free* methods, where the model parameters are only implicitly derived, if at all (Montgomery and Mastrangelo, 1991; Zhang, 1998).

The Information Theoretic Process Control (ITPC) is an example for a *model-generic* SPC method for *independent data*, proposed in (Alwan *et al*., 1998). Finally, a *model-generic* SPC method for *dependent* data is proposed in (Gal *et al*., 2003).

4. Multivariate Outlier Detection

In many cases multivariable observations can not be detected as outliers when each variable is considered independently. Outlier detection is possible only when multivariate analysis is performed, and the interactions among different variables are compared within the class of data. A simple example can be seen in Figure 1.1, which presents data points having two measures on a two-dimensional space. The lower left observation is clearly a multivariate outlier but not a univariate one. When considering each measure separately with respect to the spread of values along the x and y axes, we can is seen that they fall close to the center of the univariate distributions. Thus, the test for outliers must take into account the relationships between the two variables, which in this case appear abnormal.

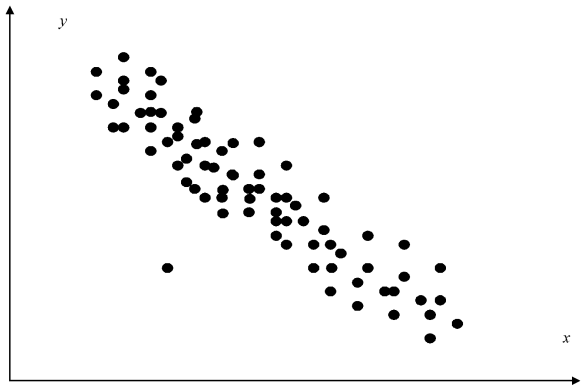

Figure 1.1. A Two-Dimensional Space with one Outlying Observation (Lower Left Corner).

Data sets with multiple outliers or clusters of outliers are subject to *masking* and *swamping* effects. Although not mathematically rigorous, the following definitions from (Acuna and Rodriguez, 2004) give an intuitive understand-

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

ing for these effects (for other definitions see (Hawkins, 1980; Iglewics and Martinez, 1982; Davies and Gather, 1993; Barnett and Lewis, 1994)):

Masking effect It is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

Swamping effect It is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers. A single step procedure with low masking and swamping is given in (Iglewics and Martinez, 1982).

4.1 Statistical Methods for Multivariate Outlier Detection

Multivariate outlier detection procedures can be divided to statistical methods that are based on estimated distribution parameters, and data-mining related methods that are typically parameter-free.

Statistical methods for multivariate outlier detection often indicate those observations that are located relatively far from the center of the data distribution. Several distance measures can be implemented for such a task. The *Mahalanobis* distance is a well-known criterion which depends on estimated parameters of the multivariate distribution. Given n observations from a p-dimensional dataset (often$n$¿¿ p), denote the sample mean vector by $\bar{\mathbf{x}}_n$ and the sample covariance matrix by $\mathbf{V}_n$ **,** where

$$\mathbf{V}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}_n\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}_n\right)^T \tag{1.3}$$

The *Mahalanobis* distance for each multivariate data point $i, i = 1, \ldots, n$, is denoted by M_i and given by

$$M_i = \left(\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}_n\right)^T \mathbf{V}_n^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}_n\right)\right)^{1/2}. \tag{1.4}$$

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Accordingly, those observations with a large *Mahalanobis* distance are indicated as outliers. Note that masking and swamping effects play an important rule in the adequacy of the Mahalanobis distance as a criterion for outlier detection. Namely, masking effects might decrease the Mahalanobis distance of an outlier. This might happen, for example, when a small cluster of outliers attracts $\bar{\mathbf{x}}_n$ and inflate $\mathbf{V}_n$ towards its direction. On the other hand, swamping effects might increase the Mahalanobis distance of non-outlying observations. For example, when a small cluster of outliers attracts $\bar{\mathbf{x}}_n$ and inflate $\mathbf{V}_n$ away from the pattern of the majority of the observations (see (Penny and Jolliffe, 2001)).

4.2 Multivariate Robust Measures

As in one-dimensional procedures, the distribution mean (measuring the location) and the variance-covariance (measuring the shape) are the two most commonly used statistics for data analysis in the presence of outliers (Rousseeuw and Leory, 1987). The use of robust estimates of the multidimensional distribution parameters can often improve the performance of the detection procedures in presence of outliers. Hadi (Hadi, 1992) addresses this problem and proposes to replace the mean vector by a vector of variable medians and to compute the covariance matrix for the subset of those observations with the smallest Mahalanobis distance. A modified version of Hadi's procedure is presented in (Penny and Jolliffe, 2001). Caussinus and Roiz (Caussinus and Roiz, 1990) propose a robust estimate for the covariance matrix, which is based on weighted observations according to their distance from the center. The authors also propose a method for a low dimensional projections of the dataset. They use the Generalized Principle Component Analysis (GPCA) to reveal those dimensions which display outliers. Other robust estimators of the location (centroid) and the shape (covariance matrix) include the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) (Rousseeuw, 1985; Rousseeuw and Leory, 1987; Acuna and Rodriguez, 2004).

4.3 Data-Mining Methods for Outlier Detection

In contrast to the above-mentioned statistical methods, data-mining related methods are often non-parametric, thus, do not assume an underlying generating model for the data. These methods are designed to manage large databases from high-dimensional spaces. We follow with a short discussion on three related classes in this category: distance-based methods, clustering methods and spatial methods.

Distance-based methods were originally proposed by Knorr and Ng (Knorr and Ng, 1997; Knorr and Ng, 1998). An observation is defined as a distance-based outlier if at least a fraction β of the observations in the dataset are further

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

than r from it. Such a definition is based on a single, global criterion determined by the parameters r and β. As pointed out in Acuna and Rodriguez (2004), such definition raises certain difficulties, such as the determination of r and the lack of a ranking for the outliers. The time complexity of the algorithm is $O(pn^2)$, where p is the number of features and n is the sample size. Hence, it is not an adequate definition to use with very large datasets. Moreover, this definition can lead to problems when the data set has both dense and sparse regions (Breunig *et al*., 2000; Ramaswamy *et al*., 2000; Papadimitriou *et al*., 2002). Alternatively, Ramaswamy et al. (Ramaswamy *et al*., 2000) suggest the following definition: given two integers v and l ($v¡ l$), outliers are defined to be the top l sorted observations having the largest distance to their v-th nearest neighbor. One shortcoming of this definition is that it only considers the distance to the v-th neighbor and ignores information about closer observations. An alternative is to define outliers as those observations having a large *average distance* to the v-th nearest neighbors. The drawback of this alternative is that it takes longer to be calculated (Acuna and Rodriguez, 2004).

Clustering based methods consider a cluster of small sizes, including the size of one observation, as clustered outliers. Some examples for such methods are the *partitioning around medoids* (PAM) and the *clustering large applications* (CLARA) (Kaufman and Rousseeuw, 1990); a modified version of the latter for spatial outliers called CLARANS (Ng and Han, 1994); and a *fractal-dimension* based method (Barbara and Chen, 2000). Note that since their main objective is clustering, these methods are not always optimized for outlier detection. In most cases, the outlier detection criteria are implicit and cannot easily be inferred from the clustering procedures (Papadimitriou *et al*., 2002).

Spatial methods are closely related to clustering methods. Lu et al. (Lu *et al*., 2003) define a spatial outlier as a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighborhood. The authors indicate that the methods of spatial statistics can be generally classified into two sub categories: *quantitative tests* and *graphic approaches*. Quantitative methods provide tests to distinguish spatial outliers from the remainder of data. Two representative approaches in this category are the Scatterplot (Haining, 1993; Luc, 1994) and the Moran scatterplot (Luc, 1995). Graphic methods are based on visualization of spatial data which highlights spatial outliers. Variogram clouds and pocket plots are two examples for these methods (Haslett *et al*., 1991; Panatier, 1996). Schiffman et al. (Schiffman *et al*., 1981) suggest using a multidimensional scaling (MDS) that represents the similarities between objects spatially, as in a map. MDS seeks to find the best configuration of the observations in a low dimensional space. Both metric and non-metric forms of MDS are proposed in (Penny and Jolliffe, 2001). As indicated above, Ng and Han (Ng and Han, 1994) develop a clustering method for spatial data-mining called CLARANS which is based on

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

randomized search. The authors suggest two spatial data-mining algorithms that use CLARANS. Shekhar et al. (Shekhar and Lu, 2001; Shekhar and Lu, 2002) introduce a method for detecting spatial outliers in graph data set. The method is based on the distribution property of the difference between an attribute value and the average attribute value of its neighbors. Shekhar et al., (Shekhar and Lu, 2003) propose a unified approach to evaluate spatial outlier-detection methods. Lu et al. (Lu *et al*., 2003) propose a suite of spatial outlier detection algorithms to minimize false detection of spatial outliers when their neighborhood contains true spatial outliers.

Applications of spatial outliers can be found in fields where spatial information plays an important role, such as, ecology, geographic information systems, transportation, climatology, location-based services, public health and public safety (Ng and Han, 1994; Shekhar and Chawla, 2002; Lu *et al*., 2003).

4.4 Preprocessing Procedures

Different paradigms were suggested to improve the efficiency of various data analysis tasks including outlier detection. One possibility is to reduce the size of the data set by assigning the variables to several representing groups. Another option is to eliminate some variables from the analyses by methods of *data reduction* (Barbara *et al*., 1996), such as methods of *principal components* and *factor analysis* that are further discussed in Chapter XXXX of this volume.

Another means to improve the accuracy and the computational tractability of multiple outlier detection methods is the use of biased sampling. Kollios et al. (Kollios *et al*., 2003) investigate the use of biased sampling according to the density of the data set to speed up the operation of general data-mining tasks, such as clustering and outlier detection.

5. Comparison of Outlier Detection Methods

Since different outlier detection algorithms are based on disjoints sets of assumption, a direct comparison between them is not always possible. In many cases, the data structure and the outlier generating mechanism on which the study is based dictate which method will outperform the others. There are few works that compare different classes of outlier detection methods.

Williams et al. (Williams *et al*., 2002), for example, suggest an outlier detection method based on *replicator neural networks* (RNNs). They provide a comparative study of RNNs with respect to two parametric (statistical) methods (one proposed in (Hadi, 1994), and the other proposed in (Knorr *et al*., 2001)) and one data-mining non-parametric method (proposed in (Oliver *et al*., 1996)). The authors find that RNNs perform adequately to the other methods in many cases, and particularly well on large datasets. Moreover, they find that some statistical outlier detection methods scale well for large dataset, despite

claims to the contrary in the data-mining literature. They summaries the study by pointing out that in outlier detection problems simple performance criteria do not easily apply.

Shekhar et al. (Shekhar and Lu, 2003) characterize the computation structure of spatial outlier detection methods and present scalable algorithms to which they also provide a cost model. The authors present some experimental evaluations of their algorithms using a traffic dataset. Their experimental results show that the *connectivity-clustered access model* (CCAM) achieves the highest clustering efficiency value with respect to a predefined performance measure. Lu et al. (Lu *et al*., 2003) compare three spatial outlier detection algorithms. Two algorithms are sequential and one algorithm based on median as a robust measure for the mean. Their experimental results confirm the effectiveness of these approaches in reducing the risk of falsely negative outliers.

Finally, Penny and Jolliffe (Penny and Jolliffe, 2001) conduct a comparison study with six multivariate outlier detection methods. The methods' properties are investigated by means of a simulation study and the results indicate that no technique is superior to all others. The authors indicate several factors that affect the efficiency of the analyzed methods. In particular, the methods depend on: whether or not the data set is multivariate normal; the dimension of the data set; the type of the outliers; the proportion of outliers in the dataset; and the outliers' degree of contamination (outlyingness). The study motivated the authors to recommend the use of a "battery of multivariate methods" on the dataset in order to detect possible outliers. We fully adopt such a recommendation and argue that the battery of methods should depend, besides on the above-mentioned factors, but also on other factors such as, the data structure dimension and size; the time constraints in regard to single vs. sequential identifiers; and whether an online or an offline outlier detection is required.

References

Acuna E., Rodriguez C. A., "Meta analysis study of outlier detection methods in classification," Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrived from academic.uprm.edu/ eacuna/paperout.pdf. In proceedings IPSI 2004, Venice, 2004.

Alwan L.C., Ebrahimi N., Soofi E.S., "Information theoretic framework for process control," European Journal of Operations Research, 111, 526-542, 1998.

Alwan L.C., Roberts H.V., "Time-series modeling for statistical process control," Journal of Business and Economics Statistics, 6 (1), 87-95, 1988.

Apley D.W., Shi J., "The GLRT for statistical process control of autocorrelated processes," IIE Transactions, 31, 1123-1134, 1999.

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Barbara D., Faloutsos C., Hellerstein J., Ioannidis Y., Jagadish H.V., Johnson T., Ng R., Poosala V., Ross K., Sevcik K.C., "The New Jersey Data Reduction Report," Data Eng. Bull., September, 1996.

Barbara D., Chen P., "Using the fractal dimension to cluster datasets," In Proc. ACM KDD 2000, 260-264, 2000.

Barnett V., Lewis T., Outliers in Statistical Data. John Wiley, 1994.

Bay S.D., Schwabacher M., "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," In Proc. of the ninth ACM-SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 2003.

Ben-Gal I., Morag G., Shmilovici A., "CSPC: A Monitoring Procedure for State Dependent Processes," Technometrics, 45(4), 293-311, 2003.

Box G. E. P., Jenkins G. M., Times Series Analysis, Forecasting and Control, Oakland, CA: Holden Day, 1976.

Breunig M.M., Kriegel H.P., Ng R.T., Sander J., "Lof: Identifying density-based local outliers," In Proc. ACMSIGMOD Conf. 2000, 93–104, 2000.

Caussinus H., Roiz A., "Interesting projections of multidimensional data by means of generalized component analysis," In Compstat 90, 121-126, Heidelberg: Physica, 1990.

David H. A., "Robust estimation in the presence of outliers," In Robustness in Statistics, eds. R. L. Launer and G.N. Wilkinson, Academic Press, New York, 61-74, 1979.

Davies L., Gather U., "The identification of multiple outliers," Journal of the American Statistical Association, 88(423), 782-792, 1993.

DuMouchel W., Schonlau M., "A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities," In Proceedings of the 4th International Conference on Knowledge Discovery and Data-mining (KDD98), 189–193, 1998.

Fawcett T., Provost F., "Adaptive fraud detection," Data-mining andKnowledge Discovery, 1(3), 291–316, 1997.

Ferguson T. S., "On the Rejection of outliers," In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 253-287, 1961.

Gather U., "Testing for multisource contamination in location / scale families," Communication in Statistics, Part A: Theory and Methods, 18, 1-34, 1989.

Grubbs F. E., "Proceadures for detecting outlying observations in Samples," Technometrics, 11, 1-21, 1969.

Hadi A. S., "Identifying multiple outliers in multivariate data," Journal of the Royal Statistical Society. Series B, 54, 761-771, 1992.

Hadi A. S., "A modification of a method for the detection of outliers in multivariate samples," Journal of the Royal Statistical Society, Series B, 56(2), 1994.

Hawkins D., Identification of Outliers, Chapman and Hall, 1980.

Hawkins S., He H. X., Williams G. J., Baxter R. A., "Outlier detection using replicator neural networks," In Proceedings of the Fifth International Conference and Data Warehousing and Knowledge Discovery (DaWaK02), Aix en Provence, France, 2002.

Haining R., Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, 1993.

Hampel F. R., "A general qualitative definition of robustness," Annals of Mathematics Statistics, 42, 1887–1896, 1971.

Hampel F. R., "The influence curve and its role in robust estimation," Journal of the American Statistical Association, 69, 382–393, 1974.

Haslett J., Brandley R., Craig P., Unwin A., Wills G., "Dynamic Graphics for Exploring Spatial Data With Application to Locating Global and Local Anomalies," The American Statistician, 45, 234–242, 1991.

Hu T., Sung S. Y., Detecting pattern-based outliers, Pattern Recognition Letters, 24, 3059-3068.

Iglewics B., Martinez J., Outlier Detection using robust measures of scale, Journal of Sattistical Computation and Simulation, 15, 285-293, 1982.

Jin W., Tung A., Han J., "Mining top-n local outliers in large databases," In Proceedings of the 7th International Conference on Knowledge Discovery and Data-mining (KDD01), San Francisco, CA, 2001.

Johnson R., Applied Multivariate Statistical Analysis. Prentice Hall, 1992.

Johnson T., Kwok I., Ng R., "Fast Computation of 2-Dimensional Depth Contours," In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, 224-228. AAAI Press, 1998.

Kaufman L., Rousseeuw P.J., Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.

Knorr E., Ng R., "A unified approach for mining outliers," In Proceedings Knowledge Discovery KDD, 219-222, 1997.

Knorr E., Ng. R., "Algorithms for mining distance-based outliers in large datasets," In Proc. 24th Int. Conf. Very Large Data Bases (VLDB), 392-403, 24-27, 1998.

Knorr, E., Ng R., Tucakov V., "Distance-based outliers: Algorithms and applications," VLDB Journal: Very Large Data Bases, 8(3-4):237-253, 2000.

Knorr E. M., Ng R. T., Zamar R. H., "Robust space transformations fordistance-based operations," In Proceedings of the 7th International Conference on Knowledge Discovery and Data-mining (KDD01), 126-135, San Francisco, CA, 2001.

Kollios G., Gunopulos D., Koudas N., Berchtold S., "Efficient biased sampling for approximate clustering and outlier detection in large data sets," IEEE Transactions on Knowledge and Data Engineering, 15 (5), 1170-1187, 2003.

Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and
Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

Liu H., Shah S., Jiang W., "On-line outlier detection and data cleaning," Computers and Chemical Engineering, 28, 1635–1647, 2004.

Lu C., Chen D., Kou Y., "Algorithms for spatial outlier detection," In Proceedings of the 3rd IEEE International Conference on Data-mining (ICDM'03), Melbourne, FL, 2003.

Lu C.W., Reynolds M.R., "EWMA Control Charts for Monitoring the Mean of Autocorrelated Processes," Journal of Quality Technology, 31 (2), 166-188, 1999.

Luc A., "Local Indicators of Spatial Association: LISA," Geographical Analysis, 27(2), 93-115, 1995.

Luc A., "Exploratory Spatial Data Analysis and Geographic Information Systems," In M. Painho, editor, New Tools for Spatial Analysis, 45-54, 1994.

Martin R. D., Thomson D. J., "Robust-resistant spectrum estimation," In Proceeding of the IEEE, 70, 1097-1115, 1982.

Montgomery D.C., Mastrangelo C.M., "Some statistical process control methods for autocorrelated data," Journal of Quality Technology, 23 (3), 179-193, 1991.

Ng R.T., Han J., Efficient and Effective Clustering Methods for Spatial Data Mining, In Proceedings of Very Large Data Bases Conference, 144-155, 1994.

Oliver J. J., Baxter R. A., Wallace C. S., "Unsupervised Learning using MML," In Proceedings of the Thirteenth International Conference (ICML96), pages 364-372, Morgan Kaufmann Publishers, San Francisco, CA, 1996.

Panatier Y., Variowin. Software for Spatial Data Analysis in 2D., Springer-Verlag, New York, 1996.

Papadimitriou S., Kitawaga H., Gibbons P.G., Faloutsos C., "LOCI: Fast Outlier Detection Using the Local Correlation Integral," Intel research Laboratory Technical report no. IRP-TR-02-09, 2002.

Penny K. I., Jolliffe I. T., "A comparison of multivariate outlier detection methods for clinical laboratory safety data," The Statistician 50(3), 295-308, 2001.

Perarson R. K., "Outliers in process modeling and identification," IEEE Transactions on Control Systems Technology, 10, 55-63, 2002.

Ramaswamy S., Rastogi R., Shim K., "Efficient algorithms for mining outliers from large data sets," In Proceedings of the ACM SIGMOD International Conference on Management of Data, Dalas, TX, 2000.

Rosner B., On the detection of many outliers, Technometrics, 17, 221-227, 1975.

Rousseeuw P., "Multivariate estimation with high breakdown point," In: W. Grossmann et al., editors, Mathematical Statistics and Applications, Vol. B, 283-297, Akademiai Kiado: Budapest, 1985.

Rousseeuw P., Leory A., Robust Regression and Outlier Detection, Wiley Series in Probability and Statistics, 1987.

Runger G., Willemain T., "Model-based and Model-free Control of Autocorrelated Processes," Journal of Quality Technology, 27 (4), 283-292, 1995.

Ruts I., Rousseeuw P., "Computing Depth Contours of Bivariate Point Clouds," In Computational Statistics and Data Analysis, 23,153-168, 1996.

Schiffman S. S., Reynolds M. L., Young F. W., Introduction to Multidimensional Scaling: Theory, Methods and Applications. New York: Academic Press, 1981.

Shekhar S., Chawla S., A Tour of Spatial Databases, Prentice Hall, 2002.

Shekhar S., Lu C. T., Zhang P., "Detecting Graph-Based Spatial Outlier: Algorithms and Applications (A Summary of Results)," In Proc. of the Seventh ACM-SIGKDD Conference on Knowledge Discovery and Data Mining, SF, CA, 2001.

Shekhar S., Lu C. T., Zhang P., "Detecting Graph-Based Spatial Outlier," Intelligent Data Analysis: An International Journal, 6(5), 451–468, 2002.

Shekhar S., Lu C. T., Zhang P., "A Unified Approach to Spatial Outliers Detection," GeoInformatica, an International Journal on Advances of Computer Science for Geographic Information System, 7(2), 2003.

Wardell D.G., Moskowitz H., Plante R.D., "Run-length distributions of special-cause control charts for correlated processes," Technometrics, 36 (1), 3–17, 1994.

Tukey J.W., Exploratory Data Analysis. Addison-Wesley, 1977.

Williams G. J., Baxter R. A., He H. X., Hawkins S., Gu L., "A Comparative Study of RNN for Outlier Detection in Data Mining," IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102, 2002.

Williams G. J., Huang Z., "Mining the knowledge mine: The hot spots methodology for mining large real world databases," In Abdul Sattar, editor, Advanced Topics in Artificial Intelligence, volume 1342 of Lecture Notes in Artificial Intelligence, 340–348, Springer, 1997.

Zhang N.F., "A Statistical Control Chart for Stationary Process Data," Technometrics, 40 (1), 24–38, 1998.